IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

GSDC Transformer: An Efficient and Effective Cue Fusion for Monocular Multi-Frame Depth Estimation

Naiyu Fang, Lemiao Qiu, Shuyou Zhang, Zili Wang, Zheyuan Zhou, and Kerui Hu

Abstract— Depth estimation provides an alternative approach for perceiving 3D information in autonomous driving. Monocular depth estimation, whether with single-frame or multi-frame inputs, has achieved significant success by learning various types of cues and specializing in either static or dynamic scenes. Recently, these cues fusion becomes an attractive topic, aiming to enable the combined cues to perform well in both types of scenes. However, adaptive cue fusion relies on attention mechanisms, where the quadratic complexity limits the granularity of cue representation. Additionally, explicit cue fusion depends on precise segmentation, which imposes a heavy burden on mask prediction. To address these issues, we propose the GSDC Transformer, an efficient and effective component for cue fusion in monocular multi-frame depth estimation. We utilize deformable attention to learn cue relationships at a fine scale, while sparse attention reduces computational requirements when granularity increases. To compensate for the precision drop in dynamic scenes, we represent scene attributes in the form of super tokens without relying on precise shapes. Within each super token attributed to dynamic scenes, we gather its relevant cues and learn local dense relationships to enhance cue fusion. Our method achieves state-of-the-art performance on the KITTI dataset with efficient fusion speed.

I. Introduction

As a step toward visual-only autonomous driving, depth estimation plays a crucial role in providing additional distance information for 3D object detection [1]–[3] and scene understanding [4], [5], potentially replacing the need for LiDAR sensors. Both single-frame [6]–[8] and multi-frame [9]–[11] monocular depth estimations have achieved remarkable success by leveraging various cues [12]. However, their estimation mechanisms have limitations in either static or dynamic scenes [13]. To circumvent their limitation, how to fuse these cues [12], [13] is blooming in the community. In this paper, we focus on this cue fusion topic, to chase an efficient and effective component that enhances monocular multi-frame depth estimation.

Monocular depth estimation provides a low-hardware-cost scheme to perceive distance. The single-frame approach learns spatial context cues like texture gradient and utilizes a deep feature extractor to estimate pixel-level depth [12]. Although significant improvements have been made through model and mechanism upgrades [14], [15], it remains an ill-posed problem, leading to artifacts in static scenes. To tackle it, inspired by stereo match [16], the multi-frame approach exploits cost volume [17] to match features between adjacent frame at different depth hypotheses. This enables significant progress in the estimation precision at static scenes by leveraging geometric cues. However, it encounters challenges in dynamic scenes due to large displacements that violate geometric consistency. There are two prevailing methods to mitigate the failure in dynamic scenes. The first method involves explicitly distinguishing dynamic scenes at the pixel level using masks [18], [19] or optical flow [20]. It replaces the estimated results or cost volume in dynamic scenes of the multi-frame approach with those from the single-frame approach and fine-tunes them at the loss level. With the emergence of Vision Transformer [21], the second method focuses on learning the spatial relationship between texture cues and geometric cues using attention mechanisms. It generates a fused result that adaptively emphasizes cues in scenes without relying

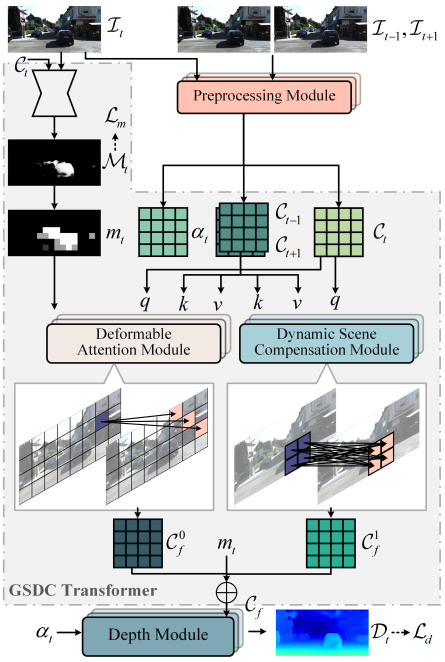

Fig. 1. The framework of monocular multi-frame depth estimation. The main focus of this paper is the cue fusion process, which is addressed by introducing a novel GSDC Transformer denoted by the gray block. GSDC Transformer is conditioned on the cost volumes of the target frame and adjacent frames, and yields a fused cost volume. It aims to increase the interaction granularity in a computation-saving manner and compensate for the precision drop in dynamic scenes.

The authors are with State Key Laboratory of Fluid Power & Mechatronic Systems, Zhejiang University, Hangzhou, 310027, China (e-mail: FangNaiyu@zju.edu.cn; qiulm@zju.edu.cn; zsy@zju.edu.cn; ziliwang@zju.edu.cn; zheyuanzhou@zju.edu.cn; hkr457@zju.edu.cn) (Corresponding authors: Lemiao Qiu)

on an explicit mask. Intrinsically, both methods aim to fuse cues to improve depth estimation performance in both static and dynamic scenes.

However, there are still some issues with the aforementioned fusion methods. 1) Due to the low pixel proportion of dynamic scenes, the detection and segmentation of pixel-level dynamic scenes require high shape accuracy, which poses a significant burden on the mask prediction module. Incorrect shapes can lead to erroneous replacements; 2) Due to the inherent quadric complexity, Transformer-based adaptive cue fusion inevitably reduces the granularity of cue representations through downsampling, impacting the final estimation precision at the pixel level, while full attention also affects fusion efficiency.

To tackle these issues, we propose a novel Globally Sparse and Dense-Compensated Transformer (GSDC Transformer), which functions as a crucial component for cue fusion in monocular multi-frame depth estimation. We strive to establish a learning paradigm for efficient and effective cue fusion, where the core idea is to fine overall granularity and compensate for local dynamic scenes. The contribution of this paper is three-fold:

(1) We propose a super dynamic scenes mask to alleviate the need for precise segmentation. It represents quantal scene attributions in the form of super token and provides enough candidates for compensating dynamic scenes, while its prediction relies on a lightweight CNN model, thus ensuring efficient fusion processes.

(2) We propose to enhance the overall fusion granularity and compensate for dynamic scenes. The deformable attention module learns fine-scale spatial relationships between cues. To mitigate the precision drop caused by sparse relationships in dynamic scenes, a dynamic compensation module supplements local dense relationships for each super token.

(3) Experiments demonstrate that our method achieves state-of-the-art performance on the KITTI dataset [22], ranking first and second for overall and dynamic scenes, respectively. Furthermore, our fusion method achieves nearly 20% savings in FLOPs compared to [13].

II. Related Work

A. Single-Frame Depth Estimation

Learning-based single-frame depth estimation has emerged with the advent of CNNs, utilizing an encoder-decoder structure to map RGB features to depth values at the pixel level. Laina et al. [6] proposed an end-to-end model for this task, eliminating the need for hand-crafted features and post-processing techniques. Due to the long-tailed distribution of depth values, direct regression of depth values encounters issues of slow convergence and local minima in end-to-end models. Several studies aimed to develop effective prediction heads and objective functions. Fu et al. [14] employed the SID strategy to convert continuous depth into discrete values. Liu et al. [7] regressed a depth probability distribution to construct a 3D depth probability volume. Jiao et al. [8] utilized attention-driven loss to establish a connection between semantic segmentation and depth estimation. Yin et al. [23] utilized geometric constraints in 3D space to penalize virtual normal directions.

Building on these frameworks and paradigms, some studies aimed to capture more global information in depth estimation. They achieved this by utilizing progressive upsampling to expand the receptive field or employing dilated convolutions [14]. With the emergence of Vision Transformers, Ranftl et al. [24] applied this task within a pure attention framework and gained a substantial precision improvement. Bhat et al. [15] further introduced global processing information and adaptively learned the range of discrete depth values based on specific instances.

B. Multi-Frame Depth Estimation

A traditional approach for multi-frame depth estimation involves capturing features across frames using a temporal model. Wang et al. [25] utilized a recurrent neural network, Patil et al. [26] employed a ConvLSTM model to learn spatiotemporal information, and Wang et al. [27] utilized a GRU-based estimator to match feature similarity in the hidden state. Drawing inspiration from multi-view stereo [28], mainstream methods construct cost volumes [9] to explore the 3D spatial relationships between multiple frames. This approach has achieved significant success in handling the ambiguity of monocular depth estimation, particularly for static scenes.

Due to the presence of moving objects in autonomous driving, such as vehicles, motorbikes, and pedestrians, and these dynamic scenes deviate from geometric assumptions, some studies endeavored to address the artifact of dynamic scenes in depth estimation. They partition dynamic scenes in calculating photometric loss [10], [11] and constructing cost volumes [19], [29], or model moving objects by leveraging the fundamental difference between inverse and forward projection [30].

The multi-view stereo also offers a paradigm for unsupervised and self-supervised learning of multi-frame depth estimation. Geometry consistency [31] serves as a fundamental component, supervising depth estimation through the projection and warping between frames. At the loss level, this constraint is further enhanced by introducing Depth Hints [32] and employing a minimum reprojection loss [33]. At the mechanism level, Shu et al. [34] proposed a feature-level projection, while Johnston et al. [35] employed self-attention to capture features and constrained the discrete disparity volume.

III. Methodology

A. Outline

Monocular multi-frame depth estimation predicts a depth map $\mathcal{D}_t$ from the target frame $\mathcal{I}_t$ at time t by learning information from adjacent frames $\mathcal{I}_{t-1}, \mathcal{I}_{t+1}$.

This paper focuses on information learning process between cues in the target frame and adjacent frames, i.e. cue fusion. As Fig. 1 shows, in the preprocessing module, we exploit the depth converting method [13] to obtain the pseudo cost volume $\mathcal{C}_t \in \mathbb{R}^{H\times W\times D}$ of $\mathcal{I}_t$, where D represents the number of discrete depth channels. We utilize SSIM-based photometric error method [18] to construct cost volumes $\mathcal{C}_{t-1}, \mathcal{C}_{t+1} \in \mathbb{R}^{H\times W\times D}$ conditioned on $\mathcal{I}_{t-1}, \mathcal{I}_t, \mathcal{I}_{t+1}$, and these cost volumes are concatenated at depth channel instead of weighted summation. Additionally, we employ a backbone (ResNet18 [36] or Efficient-B5 [37]) to extract the multi-scale features α_t from $\mathcal{I}_t$. Based on these inputs, we propose a novel GSDC Transformer to combine cues from adjacent frame into the target frame, resulting in the prediction of a fused cost volume $\mathcal{C}_f \in \mathbb{R}^{H\times W\times D}$. Eventually, $\mathcal{C}_f$ is fed into a depth module [13] to predict the depth map $\mathcal{D}_t \in \mathbb{R}^{H\times W\times 1}$ of the target frame.

To achieve efficient and effective multi-frame fusion, we devise GSDC Transformer to consist of a deformable attention module, a super dynamic scene mask prediction module, and a dynamic scene compensation module. Inspired by Deformable DETR [38], the deformable attention module (in Sec. III-B.1) establishes sparse attention relationships between $\mathcal{C}_t$ and $\mathcal{C}_{t-1}, \mathcal{C}_{t+1}$ for each token. By employing linear complexity instead of quadratic complexity, the deformable attention module can be implemented at a finer grain, enhancing the precision of overall scenes. To compensate for the precision drop in dynamic scenes, we adopt a lightweight CNN module to partition dynamic scenes (in Sec. III-B.2) into super tokens without relying on precise shape prediction. Finally, the dynamic scene compensation module (in Sec. III-B.3) focuses on learning local and dense attention relationships exclusively within dynamic scenes to facilitate its cue fusion.

B. GSDC Transformer

1) Fine-Grained and Sparse Attention but Precision Drop in Dynamic Scenes: According to [12], monocular single-frame depth estimation relies on texture spatial context cues such as texture gradient, while monocular multi-frame depth estimation focuses on learning geometric cues between adjacent frames. Intrinsically, we aim to fuse the geometric cues from $\mathcal{C}_{t-1}, \mathcal{C}_{t+1}$ with spatial context cues in $\mathcal{C}_t$ by learning their correspondence relationship, incorporating both types of cues in the fused cost volume $\mathcal{C}_f$. A previous study [13] employed full attention to learn the relationships between these cues. However, the quadratic complexity of attention led to downsampling the cost volumes by a factor of 4× to manage computational costs. Unfortunately, this downsampling results in a coarse-grained token representation, leading to a precision drop in cue fusion.

We aim to increase the granularity of cue representation while maintaining computational efficiency. To achieve this, a straightforward and well-known approach is to employ sparse attention instead of full attention. Since cost volumes are constructed in a pixel-aligned manner and most pixels exhibit minimal movement between adjacent frames, there is an opportunity to explore the correspondence relationship in the cue fusion via a sparse manner. Therefore, in the deformable attention module, we increase the overall cue granularity by downsampling tokens by a factor of 2×, and the attention is sparse to ensure an efficient cue fusion. Experimental results presented in Table II demonstrate that sparse attention does not significantly compromise overall precision. Specifically, following the paradigm of deformable DETR [38], as shown in Fig. 2a, we utilize $\mathcal{C}_t$ as the query to predict weights $w \in \mathbb{R}^{N_q\times N_s\times 1}$ and offsets $p \in \mathbb{R}^{N_q\times N_s\times 1}$, where N_s represents the number of sample points with $N_s \ll N_v$. The concatenated $\mathcal{C}_{t-1}, \mathcal{C}_{t+1}$ serve as the value, and N_s points are sampled for each token, resulting in $v_s \in \mathbb{R}^{N_q\times N_s\times d_v}$. Finally, the product of v_s and w yields the predicted fused cost volume $\mathcal{C}_f^0$ through projection and upsampling.

However, it is crucial to consider dynamic scenes despite their low pixel percentage [12]. Due to the noticeable positional shifts of moving objects between adjacent frames, more correspondence relationships need to be learned in cue fusion by maintaining the preservation and enhancement of spatial context cues. Unfortunately, as shown in Table II, the sparse attention fails to effectively handle dynamic scenes, leading to a significant precision drop.

2) Super Dynamic Scene Mask: To maintain our desired efficiency, a straightforward approach is to supplement full attention only to dynamic scenes. However, this approach heavily relies on an explicit semantic segmentation of dynamic scenes, necessitating a precise dynamic scene mask. In this section, our goal is to mitigate the reliance on this precise shape prior. Inspired by the window local attention mechanism [39], as shown in Fig. 3, we exploit windows to partition the dynamic scene mask $\mathcal{M}_t$ into a super dynamic scene mask m_t, where $\kappa \times \kappa$ tokens are grouped within each window to form a super token. Each super token is assigned a unified scene attribution as Equ. (1).

$$m_t(x,y) = \max\left\{\mathcal{M}_t(u,v) \,|u \in [x\kappa, x\kappa+\kappa), v \in [y\kappa, y\kappa+\kappa)\right\} \tag{1}$$

The super dynamic scene mask is designed to provide a quantal attribution for each super token, rather than representing the shape at the pixel level. This converting approach also reduces the prediction burden of the model. Specifically, we concatenate the target frame $\mathcal{I}_t$ and the pseudo cost volume $\mathcal{C}_t$, fed them into a lightweight U-Net [40] to predict $\mathcal{M}_t$, and convert it into m_t. As shown in Fig. 3, although the predicted dynamic scene mask may exhibit shape differences compared to the ground truth, it still provides an adequate number of super token candidates attributed to dynamic scenes.

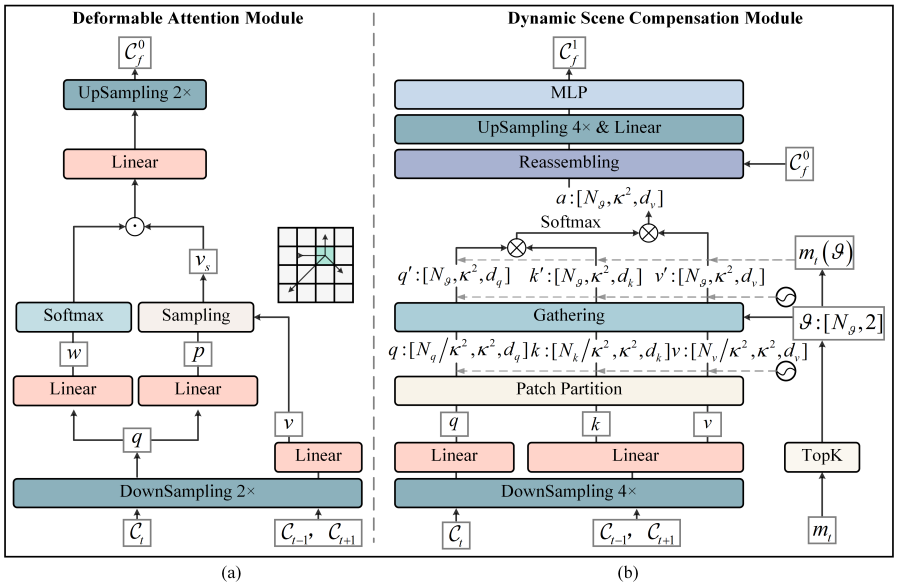

Fig. 2. The detailed structures of the deformable attention module and the dynamic scene compensation module. (a) The deformable attention module increases the granularity of cue representation with only downsampling 2× and employs sparse attention to save computation costs. (b) The dynamic scene compensation module gathers tokens attributed to super dynamic scene tokens and learns their local dense relationships to compensate for the precision drop.

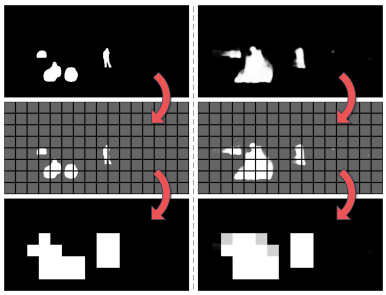

Fig. 3. The left and right parts represent the super dynamic scene masks of the predicted result and the ground truth, respectively. Compared to the original form, the super token form can represent dynamic scenes without relying on the precise shape.

3) Compensate for Dynamic Scenes: The super dynamic scene mask enables us to establish connections between each super token attributed to dynamic scenes and a local window in the Swin Transformer [39], allowing for full attention among tokens within each super token. To achieve this, we propose a dynamic compensation module as shown in Fig. 2b, which takes $\mathcal{C}_t$ as the query, and $\mathcal{C}_{t-1}, \mathcal{C}_{t+1}$ as the key and value, downsampling them by a factor of 4×. We partition q, k, v using windows of size $\kappa \times \kappa$, and embed a learned position matrix at N_q/κ^2 dimension.

Since the number of super tokens attributed to dynamic scenes varies across instances, and the standard Swin Transformer performs calculations in windows-parallel, it is important to maintain stability during training and inference. To address this, we employ a TopK function to extract N_ϑ ($N_\vartheta \ll (HW/\kappa^2)$) candidate as Equ. (2), where ϑ is the index matrix. Based on ϑ, q', k', v' are gathered from q, k, v with a relative position embedding within each window.

$$m_t(\vartheta), \vartheta = TopK(m_t) \tag{2}$$

In some cases, the tail parts of candidate tokens may have small scene values, indicating that they do not correspond to real dynamic scenes. To address this, we multiply q', k', v' with $m_t(\vartheta)$ to encourage the model to focus on real dynamic scene parts. Subsequently, we learn $a \in \mathbb{R}^{N_\vartheta \times \kappa^2 \times d_v}$ using the well-known local window attention mechanism. In order to restore the original spatial structure, we reassemble a using the base $\mathcal{C}_f^0$ and the index ϑ. Through upsampling and MLP, we predict another fused cost volume $\mathcal{C}_f^1$.

Finally, we fuse $\mathcal{C}_f^0$ and $\mathcal{C}_f^1$ using weighted summation and a conv 1×1 layer as Equ. (3).

$$\mathcal{C}_f = \text{Conv}\left[\mathcal{C}_f^0 \cdot (1 - m_t) + \mathcal{C}_f^1 \cdot m_t\right] \tag{3}$$

4) Objective Function: The objective function of super dynamic mask prediction module is formulated as Equ. (4) where BCE$(\cdot)$ is the binary entropy loss, and $\tilde{\mathcal{M}}_t$ is the ground truth. The objective function of the depth module remains the same as [13], including the scale-invariant loss [15], virtual normal loss [41], and a monocular depth loss.

$$\ell_m = \text{BCE}\left(\mathcal{M}_t, \tilde{\mathcal{M}}_t\right) \tag{4}$$

IV. Experiment

A. Implementation Details

TABLE I

The quantitative comparisons between different methods with the same frame size 256×512. BB and SV represent backbone and supervision. Bold and underline highlights represent the best and second-best performance.

Overall Scenes	BB	SV	Abs Rel	Sq Rel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ours	Res-18	Full	0.040	0.135	1.994	0.068	0.980	0.996	0.999
Manydepth [29]	Res-18	Self	0.071	0.343	3.184	0.108	0.945	0.991	0.998
DynamicDepth [19]	Res-18	Self	0.068	0.296	3.067	0.106	0.945	0.991	0.998
MonoRec [18]	Res-18	Semi	0.050	0.290	2.266	0.082	0.972	0.991	0.996
DMDepth [13]	Res-18	Full	0.043	0.151	2.113	0.073	0.975	0.996	0.999
Ours	Eff-b5	Full	0.040	0.109	1.815	0.064	0.984	0.998	0.999
DMDepth [13]	Eff-b5	Full	0.046	0.155	2.112	0.076	0.973	0.996	0.999
MaGNet [12]	Eff-b5	Full	0.057	0.215	2.597	0.088	0.967	0.996	0.999

Dynamic Scenes	BB	SV	Abs Rel	Sq Rel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ours	Res-18	Full	0.163	1.484	5.309	0.188	0.791	0.930	0.974
Manydepth [29]	Res-18	Self	0.222	3.390	7.921	0.237	0.676	0.902	0.964
DynamicDepth [19]	Res-18	Self	0.208	2.757	7.362	0.227	0.682	0.911	0.971
MonoRec [18]	Res-18	Semi	0.360	9.083	10.963	0.346	0.590	0.882	0.780
DMDepth [13]	Res-18	Full	0.118	0.835	4.297	0.146	0.871	0.975	0.990
Ours	Eff-b5	Full	0.135	1.002	4.774	0.160	0.829	0.968	0.992
DMDepth [13]	Eff-b5	Full	0.111	0.768	4.117	0.135	0.881	0.980	0.994
MaGNet [12]	Eff-b5	Full	0.141	1.219	4.877	0.168	0.830	0.955	0.986

1) Dataset: We use the Odometry version [18] of the KITTI dataset [22], which consists of 13,666 samples in the training dataset and 8,634 samples in the testing dataset. The frames are cropped and resized to 256×512, and the depth range is set from 0 to 80 m. Furthermore, we utilized the estimated pose [42] and the depth ground truth [43] to supervise the final depth estimation, and we leveraged the mask ground truth [18] to supervise the training of the super dynamic scene mask prediction module.

2) Training: All experiments are conducted on a single NVIDIA GPU 3090. To ensure comprehensive coverage of the dynamic scene compensation module, we first train the super dynamic scene mask prediction module and then train the other components based on the predicted mask prior. When training the super dynamic scene mask prediction module, we set the batch size to 32 and train for 80 epochs. We employ the Adam optimizer with the oncyclelr learning strategy [44], where the maximum learning rate is set to 2e-3 and the start and end division factors are both 10. It is important to note that, since dynamic scenes have a low pixel proportion and may not appear in some instances, during training, the IoU initially increases and then rapidly decreases, indicating a pattern collapse. As described in Sec. III-B.2, we do not require a precise shape prediction. Therefore, we select the best-trained model with the highest covering proportion (CP) as Equ. (5), which provides sufficient candidates for the dynamic scene compensation module. Additionally, the training settings of the other components are the same as in [13].

$$CP = \frac{TopK\left(m_t\right) \cap \tilde{m}_t}{\tilde{m}_t} \tag{5}$$

B. Comparison

We compared our method with Manydepth [29], DynamicDepth [19], MonoRec [18], MaGNet [12], and DMDepth [13]. The quantitative and qualitative results are shown in Table I and Fig. 4. In overall scenes, our method outperforms all other methods in terms of all evaluated metrics. Compared to second best method DMDepth, our method decreases Abs Rel, Sq Rel, RMSE by 6.98%, 10.57%, 5.63%, respectively. In dynamic scenes, our method ranks second, following DMDepth, but still significantly outperforms the other methods. Moreover, the cue fusion within our method exhibits a 3.24 G FLOPs and achieves a 227.41 FPS, surpassing the performance of DMDepth (3.82 G FLOPs and 171.73 FPS). A detailed efficiency comparison will be further discussed in Sec. IV-C.

C. Ablation Study

1) Attention in Cue Fusion: To assess the efficiency and precision of various attention mechanisms used in cue fusion, we conducted four comparison groups. For a fair comparison, the FLOPs of our case only contain the deformable attention module and the dynamic scene compensation module. The second group (Deform) employed sparse deformable attention alone [38] with a downsampling factor of 2×. The third group (Local) utilized local window attention [39] for overall patches with a downsampling factor of 4×. The fourth group (full) employed full attention as described in [13] with a downsampling factor of 4×. The fifth group (Swin+Local) replaces the deformable attention module as a one-stage Swin transformer with a downsampling factor of 2×, and employs the remaining two modules to compensate for dynamic scenes. In Table II, when

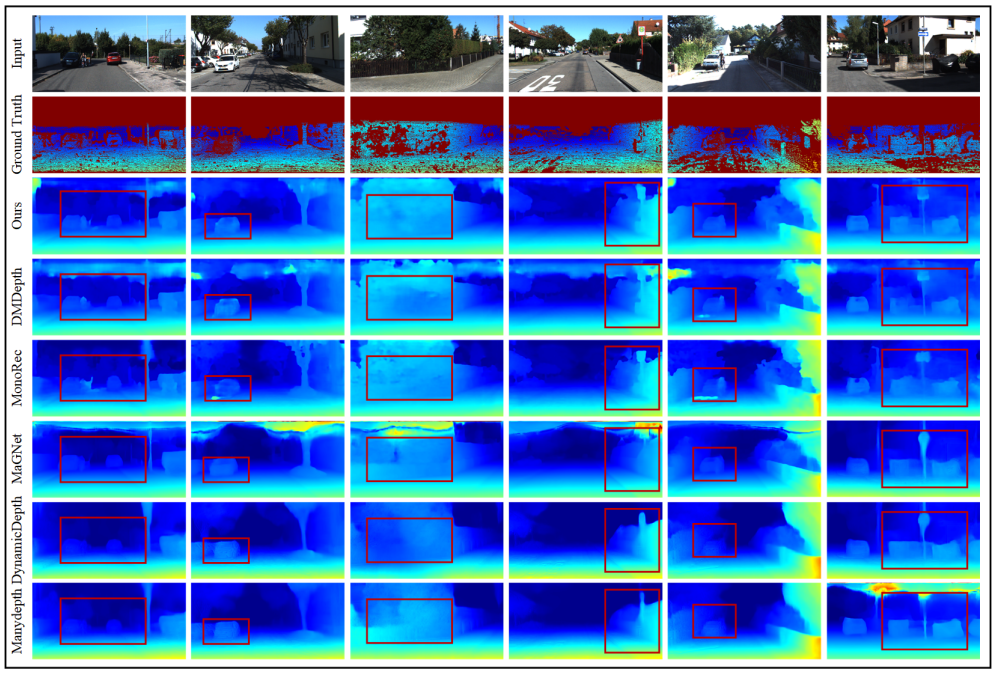

Fig. 4. The qualitative comparisons between different methods on the KITTI dataset.

considering overall scenes, deformable attention enhances precision as it refines cue granularity, while maintaining high cue fusion efficiency. Regarding dynamic scenes, deformable attention exhibits a significant decrease in precision compared to full attention due to its sparse relationship learning, as discussed in Sec. III-B.1. In addition, the superior performance of local window attention compared to deformable attention demonstrates that local dense relationship learning enhances spatial exploration among different cues. The fifth group enhances the local relationship learning between cues at the fine granularity, while omitting global relationship exploration. This combination of attention mechanisms contributes to improved performance specifically in dynamic scenes, as opposed to overall scenes. It is worth noting that that deploying a one-stage Swin transformer on fine granularity results in a significant increase in FLOPs, reaching 7.21 G, thereby contradicting the high-efficiency objective. In contrast, we utilize a super dynamic scene mask to mediate between deformable and local window attentions, adapting them to different scenes and achieving a trade-off of precision between overall and dynamic scenes. Nonetheless, the computational cost incurred by attention in our cue fusion is a mere 2.84 G FLOPs. Even with the incorporation of super dynamic scene mask prediction, the total cost for our cue fusion can remain at 3.24 G FLOPs, resulting in a 15.2% cost reduction compared to full attention.

2) Lightweight of Mask Prediction: As GSDC Transformer introduces a super dynamic scene mask prediction module to guide the compensation, in this section, we describe its lightweight design without compromising

TABLE II

The application of different types of attention in cue fusion. DS represents the downsampling factor. Cyan and Purple highlights denote the overall and dynamic scenes performances, respectively. Bold and underline highlights represent the best and second-best performance.

	DS	Abs Rel	$\delta < 1.25$	Abs Rel	$\delta < 1.25$	FLOPs/G	FPS
Ours	/	<u>0.040</u>	<u>0.980</u>	0.163	0.791	2.84	305.51
Deform	2×	0.039	0.983	0.265	0.602	1.03	834.46
Local	4×	0.079	0.919	0.200	0.694	<u>1.69</u>	<u>510.11</u>
Full	4×	0.043	0.975	0.118	0.871	3.82	171.73
Swin+Local	/	0.045	0.975	<u>0.157</u>	<u>0.798</u>	7.21	131.76

fusion efficiency. Since the $TopK(\cdot)$ operation in Equ. (2) corresponds to the one shown in Fig. 2b, we evaluate CP metric for $N_\vartheta = 6$ and $N_\vartheta = 12$. As shown in Table III, row 2 represents no downsampling and upsampling in the prediction of $\mathcal{M}_t$, Despite the relatively low shape precision (IoU), it still enables m_t to provide candidates to the dynamic scene compensation module. Since the downsampling scale of the dynamic compensation module is 4×, we explore a lightweight prediction approach in row 3 and 4, which directly performs downsampling and upsampling during the prediction of $\mathcal{M}_t$. Experimental results indicate that this coarse-scale prediction does not cause a significant drop in CP. Moreover, the FLOPs for row 3 amount to only 6.3% of those in row 2, while the FLOPs for row 4 are reduced to just 0.40 G. Another alternative method in row 5 employs a MaxPooling layer with a kernel size of $\kappa \times \kappa$ at the prediction head, and it also achieves good performance.

3) Parameter Setting: For efficient fusion, we set the default values of N_s=4 and N_ϑ=12. Upon counting

TABLE III

The lightweight design of the super dynamic scene mask prediction module. DS, US, and MP represent downsampling, upsampling, and MaxPooling.

	IoU	$CP_{N_\vartheta=6}$	$CP_{N_\vartheta=12}$	FLOPs/G
No DS & US	0.065	0.946	0.970	25.37
DS & US (4×)	0.053	0.944	0.971	1.59
DS & US (8×)	0.046	0.936	0.963	0.40
DS & MP (4×)	0.049	0.939	0.967	1.59

all samples in the dataset [22], we found that within the range of $\mu \pm 2\sigma$ in the normal distribution, the proportion of super tokens attribute to dynamic scenes to all super tokens is approximately 4%. However, with 6 and 12 candidates in our setting, the proportions increase to 5% and 10%, respectively, indicating that they cover almost all dynamic scenes in all samples. Consequently, As shown in Table IV, row 2 and row 3 exhibit similar performances, and further increasing N_ϑ to 25 does not lead to a further enhancement. Furthermore, increasing N_s to 8 results in improved overall performances at the expense of computational cost.

TABLE IV

The quantitative comparison under different parameter settings. Cyan and Purple highlights denote the overall and dynamic scenes performances, respectively.

N_s	N_ϑ	Abs Rel	$\delta < 1.25$	Abs Rel	$\delta < 1.25$	FLOPs/G	FPS
4	12	0.040	0.980	0.163	0.791	2.84	305.51
4	6	0.039	0.981	0.166	0.768	2.84	313.39
4	25	0.038	0.983	0.170	0.772	2.84	299.41
8	12	0.037	0.983	0.175	0.772	2.89	279.16

V. Conclusion

This paper proposes the GSDC Transformer as an efficient and effective component for cue fusion in monocular multi-frame depth estimation. The primary objective of this paper is to introduce a design paradigm about granularity and sparseness in cue fusion, considering both overall and dynamic scenes. We enhance the granularity of cue representation and utilize deformable attention to learn their relationships at a fine scale, resulting in a significant improvement in overall precision. However, due to the significant displacement of moving objects between frames, sparse attention is inadequate for exploring the relationships between their cues, resulting in a precision drop. To address this issue, we first represent scene attributes in the form of super tokens without relying on precise shapes. Subsequently, we select a specific number of candidates using the TopK function. Within each super token attribute to dynamic scenes, the relevant cues are gathered, and local dense relationships are learned to enhance cue fusion. Experiments demonstrate that our method achieves state-of-the-art performance on the KITTI dataset, ranking first and second for overall and dynamic scenes, respectively, and our fusion method achieves nearly a 20% reduction in FLOPs.

While our method demonstrates competitive performance, its precision in dynamic scenes falls short of certain state-of-the-art methods. In future work, we will continue to refine the compensation mechanism and framework to bolster performance specifically in dynamic scenes. Notably, the training of the super dynamic scene mask prediction module needs ground truth for dynamic scenes, incurring additional costs for manual labeling. Consequently, we also aim to devise an adaptive compensation to mitigate the reliance on this manual labeling.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No.52375271); the Natural Science Foundation of Zhejiang Province (No.LY23E050011).

References

[1] Z. Zhou, L. Du, X. Ye, Z. Zou, X. Tan, L. Zhang, X. Xue, and J. Feng, "Sgm3d: Stereo guided monocular 3d object detection," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 10 478–10 485, 2022.

[2] Y. Su, Y. Di, G. Zhai, F. Manhardt, J. Rambach, B. Busam, D. Stricker, and F. Tombari, "Opa-3d: Occlusion-aware pixel-wise aggregation for monocular 3d object detection," IEEE Robotics and Automation Letters, vol. 8, no. 3, pp. 1327–1334, 2023.

[3] I. Mouawad, N. Brasch, F. Manhardt, F. Tombari, and F. Odone, "Time-to-label: Temporal consistency for self-supervised monocular 3d object detection," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 8988–8995, 2022.

[4] G. Humblot-Renaux, L. Marchegiani, T. B. Moeslund, and R. Gade, "Navigation-oriented scene understanding for robotic autonomy: learning to segment driveability in egocentric images," IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 2913–2920, 2022.

[5] Y. B. Can, A. Liniger, O. Unal, D. Paudel, and L. Van Gool, "Understanding bird's-eye view of road semantics using an onboard camera," IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 3302–3309, 2022.

[6] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in 2016 Fourth international conference on 3D vision (3DV), 2016, pp. 239–248.

[7] C. Liu, J. Gu, K. Kim, S. G. Narasimhan, and J. Kautz, "Neural rgb (r) d sensing: Depth and uncertainty from a video camera," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10 986–10 995.

[8] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 53–69.

[9] Z. Mi, C. Di, and D. Xu, "Generalized binary search network for highly-efficient multi-view stereo," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12 991–13 000.

[10] R. Ranftl, V. Vineet, Q. Chen, and V. Koltun, "Dense monocular depth estimation in complex dynamic scenes," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4058–4066.

[11] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 582–600.

[12] G. Bae, I. Budvytis, and R. Cipolla, "Multi-view depth estimation by fusing single-view depth probability with multi-view geometry," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2842–2851.

[13] R. Li, D. Gong, W. Yin, H. Chen, Y. Zhu, K. Wang, X. Chen, J. Sun, and Y. Zhang, "Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21 539–21 548.

[14] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2002–2011.

[15] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4009–4018.

[16] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2495–2504.

[17] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2011, pp. 2320–2327.

[18] F. Wimbauer, N. Yang, L. Von Stumberg, N. Zeller, and D. Cremers, "Monorec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6112–6122.

[19] Z. Feng, L. Yang, L. Jing, H. Wang, Y. Tian, and B. Li, "Disentangling object motion and occlusion for unsupervised multi-frame monocular depth," in Proceedings of the European Conference on Computer Vision (ECCV), 2022, pp. 228–244.

[20] X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, and J. Kopf, "Consistent video depth estimation," ACM Transactions on Graphics (ToG), vol. 39, no. 4, pp. 71–1, 2020.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations (ICLR), 2021.

[22] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in 2012 IEEE conference on computer vision and pattern recognition, 2012, pp. 3354–3361.

[23] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5684–5693.

[24] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12 179–12 188.

[25] R. Wang, S. M. Pizer, and J.-M. Frahm, "Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5555–5564.

[26] V. Patil, W. Van Gansbeke, D. Dai, and L. Van Gool, "Don't forget the past: Recurrent depth estimation from monocular video," IEEE Robotics and Automation Letters, vol. 5, no. 4, pp. 6813–6820, 2020.

[27] F. Wang, S. Galliani, C. Vogel, and M. Pollefeys, "Itermvs: Iterative probability estimation for efficient multi-view stereo," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8606–8615.

[28] S. Cheng, Z. Xu, S. Zhu, Z. Li, L. E. Li, R. Ramamoorthi, and H. Su, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2524–2534.

[29] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1164–1174.

[30] S. Lee, S. Im, S. Lin, and I. S. Kweon, "Learning monocular depth in dynamic scenes via instance-aware projection consistency," in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 35, no. 3, 2021, pp. 1863–1872.

[31] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.

[32] J. Watson, M. Firman, G. J. Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2162–2171.

[33] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3828–3838.

[34] C. Shu, K. Yu, Z. Duan, and K. Yang, "Feature-metric loss for self-supervised learning of depth and egomotion," in Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 572–588.

[35] A. Johnston and G. Carneiro, "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4756–4765.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[37] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning(ICML), 2019, pp. 6105–6114.

[38] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in International Conference on Learning Representations (ICLR), 2021.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10 012–10 022.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 2015, pp. 234–241.

[41] W. Yin, Y. Liu, and C. Shen, "Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 7282–7295, 2021.

[42] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 817–833.

[43] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in 2017 international conference on 3D Vision (3DV), 2017, pp. 11–20.

[44] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, vol. 11006, 2019, pp. 369–386.